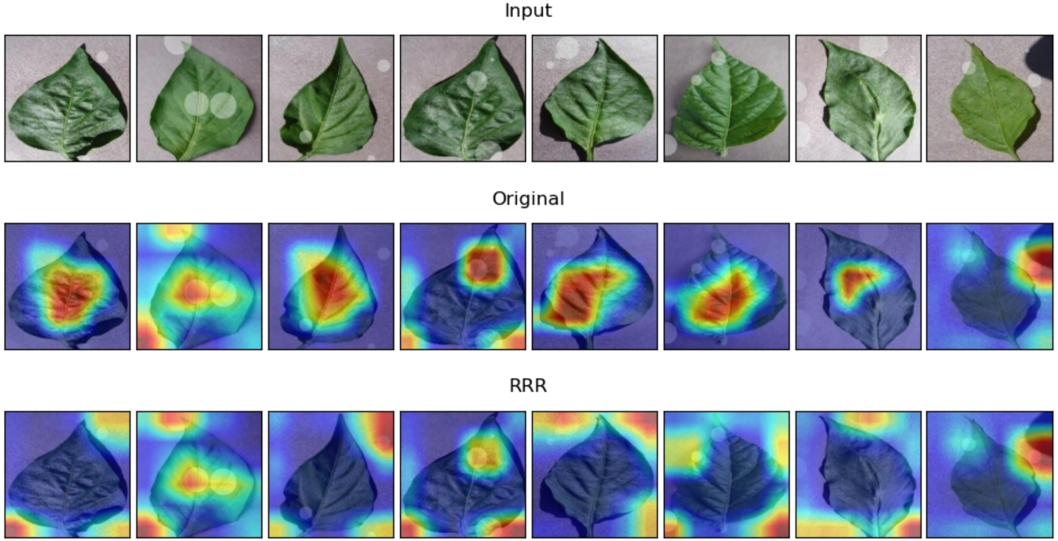
IMPROVE THE DEEP LEARNING MODELS BASED ON EXPLANATIONS AND EXPERTISE

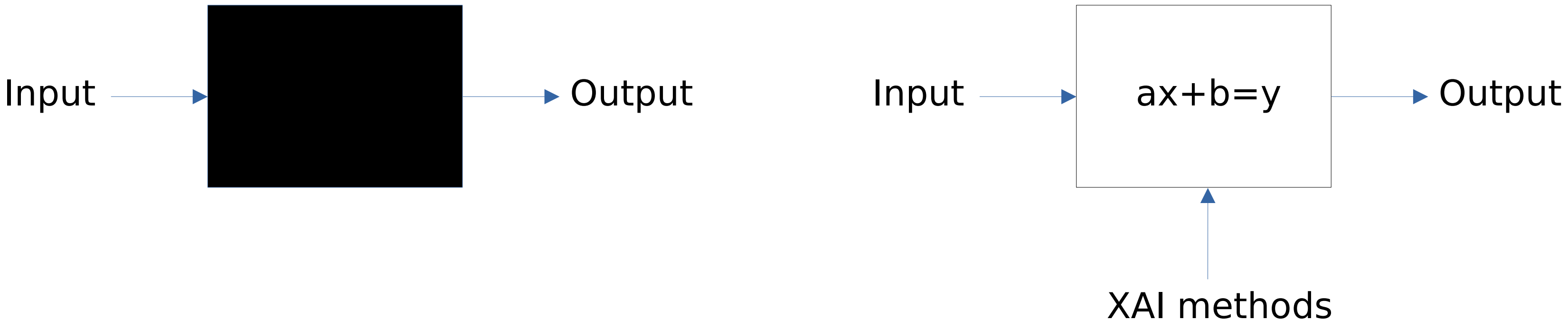


Dr. Ximeng Cheng

Fraunhofer HHI

Background

- Deep learning models have been used in many fields
- Unclear model decisions (i.e., Black-box) undermine the credibility of the results
- Many explainable artificial intelligence (XAI) methods are proposed to explain the model decisions (i.e., open the black-box)

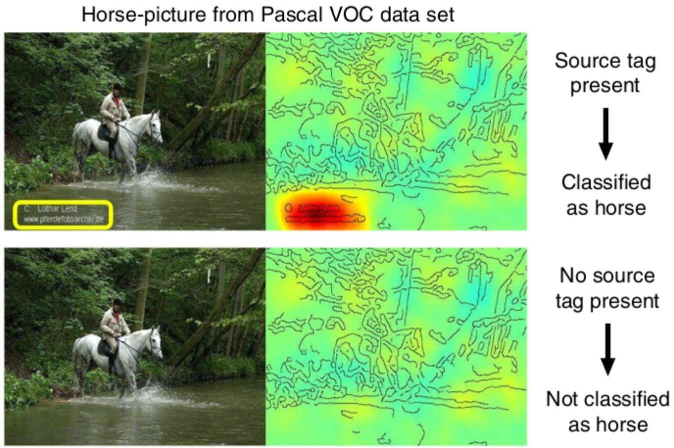


Background

- Visualization methods
  - Clustering the original model parameters
  - Displaying feature maps of part layers
- Model-agnostic methods
  - Individual conditional expectation (ICE)
  - Local interpretable model-agnostic explanations (LIME)
  - Shapley additive explanations (SHAP)
- Deep-learning-specific methods
  - Layer-wise relevance propagation (LRP)
  - Gradient-weighted class activation mapping (Grad-CAM)

Background

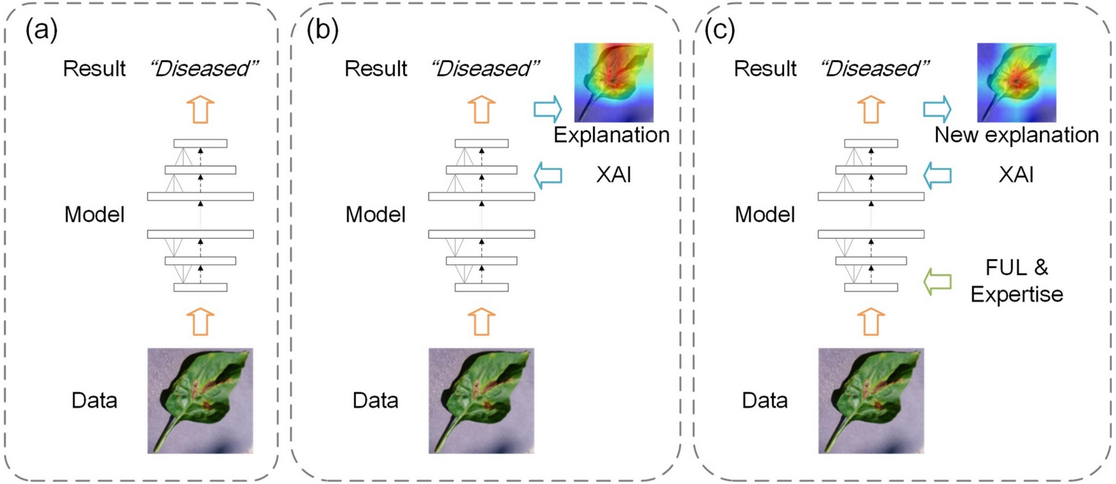
- Wrong decisions: Clever Hans effect
- Feature unlearning (FUL) methods
  - Guide the training of deep learning models
- Common FUL methods
  - Change train data:
  - e.g., Explanatory interactive learning (XIL)
  - Design new loss function:
  - e.g., Right for the right reasons (RRR)



S. Lapuschkin et al., 2019

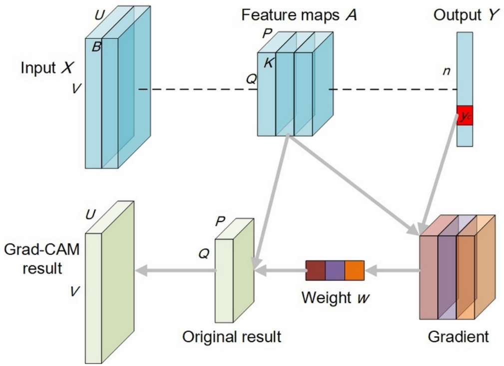
Research framework

- Three research frameworks of deep learning studies
  - Get the results only
  - Get the results and model explanations
  - Improve the model based on explanations and expertise



Methods

- Train an original model
- Use XAI methods to get the model explanations
  - XAI in this study: Grad-CAM
- Use FUL methods to introduce the expertise if current explanations are wrong
  - FUL in this study: RRR
- Retrain the model with the introduced expertise



$$Grad_{X_i} = \frac{\partial \log_e (Y_i + 1)}{\partial X_i}$$

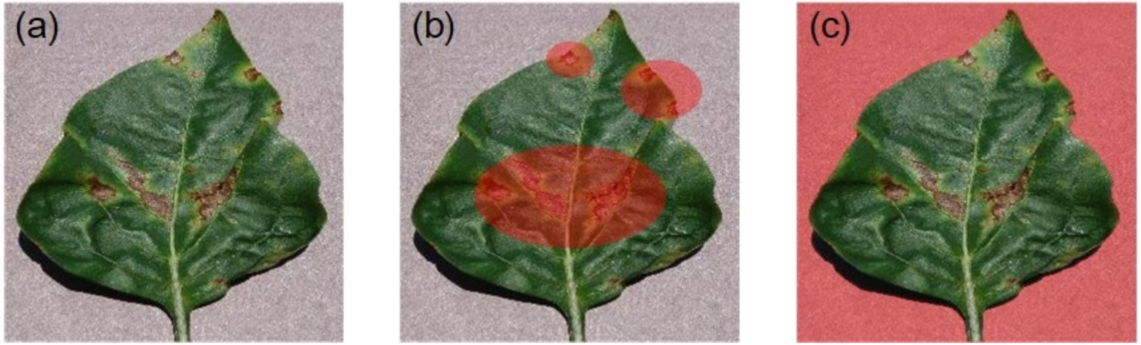
$$RLoss(X_i, Y_i, \theta, A_i) = Sum(A_i \cdot Grad_{X_i})$$

$$NLoss = CLoss + \lambda \cdot Balance(RLoss, CLoss)$$

$$Balance(l_1, l_2) = 10^{\lceil \log_{10} (\frac{l_2}{l_1}) \rceil} \cdot l_1,$$

Experimental data

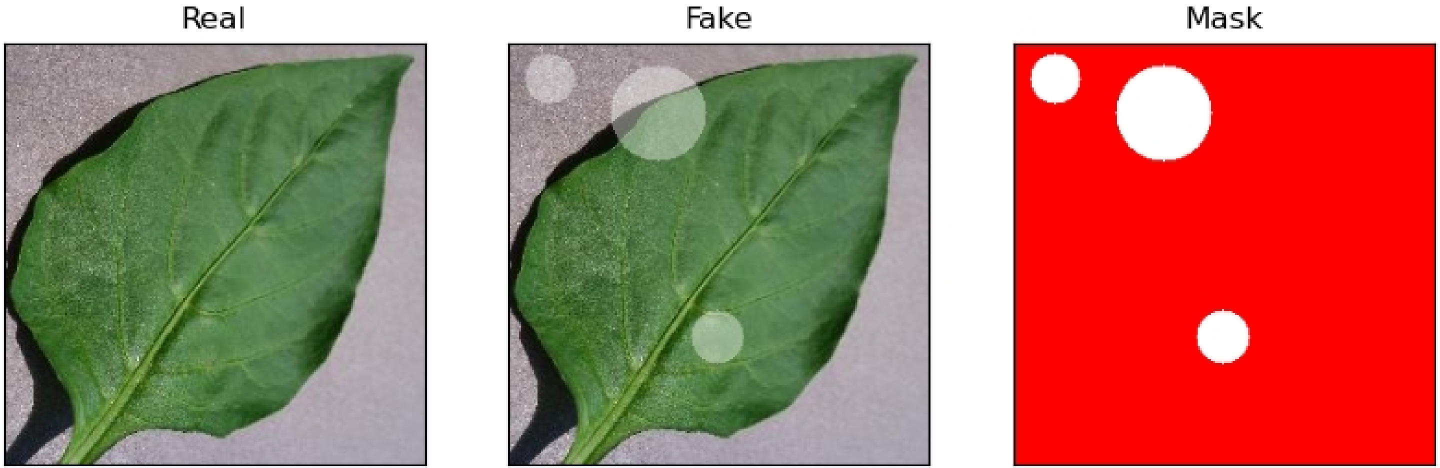
- Open-source PlantVillage dataset (https://github.com/spMohanty/PlantVillage-Dataset)
- Leaf images of multiple plant species
- With labels such as healthy and diseased



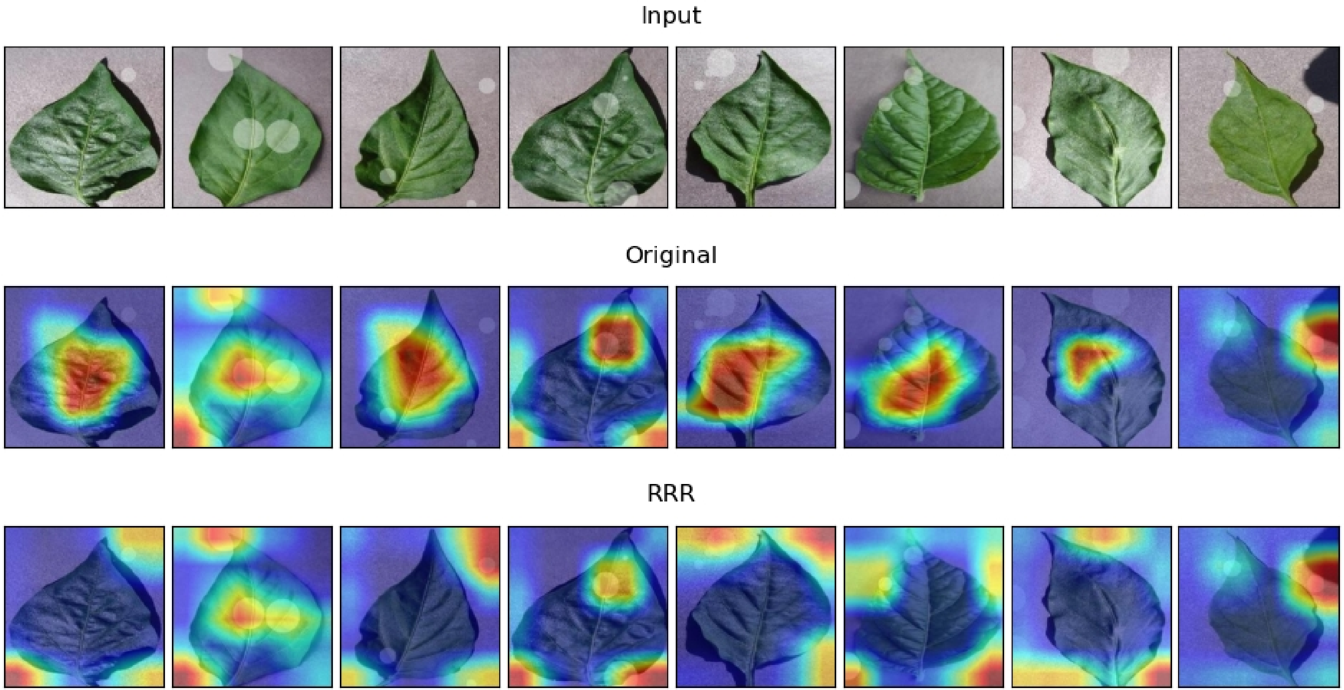
Generate the real masks of each sample

#1 Distinguishing between real and fake data

- Fake data: randomly add some transparent circles into the healthy leaf
- Real mask (pixels are useless for this task): pixels outside the added circles



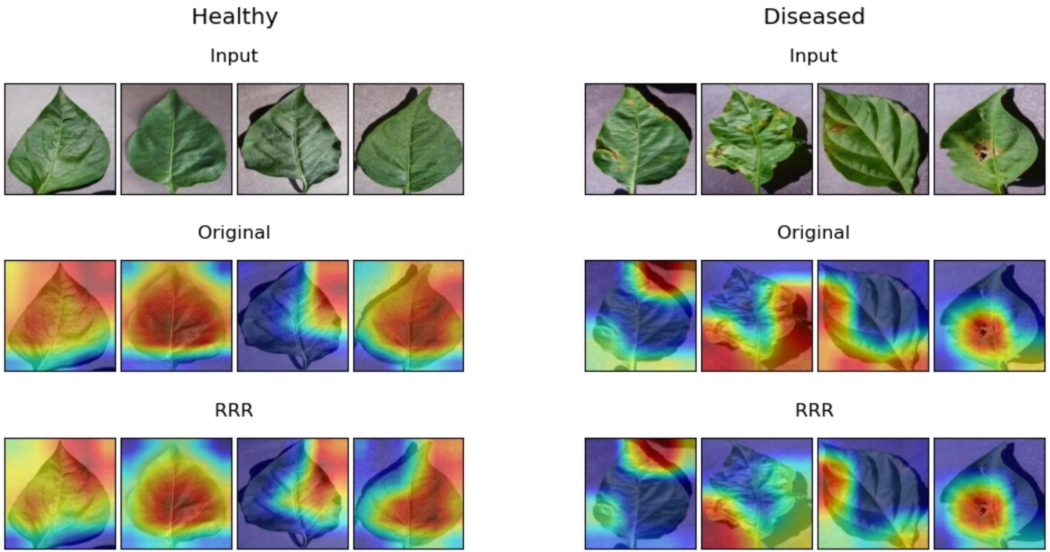
#1 Distinguishing between real and fake data



Ximeng Cheng

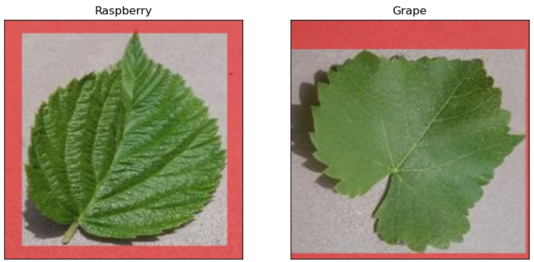
#2 Identifying diseased leaves

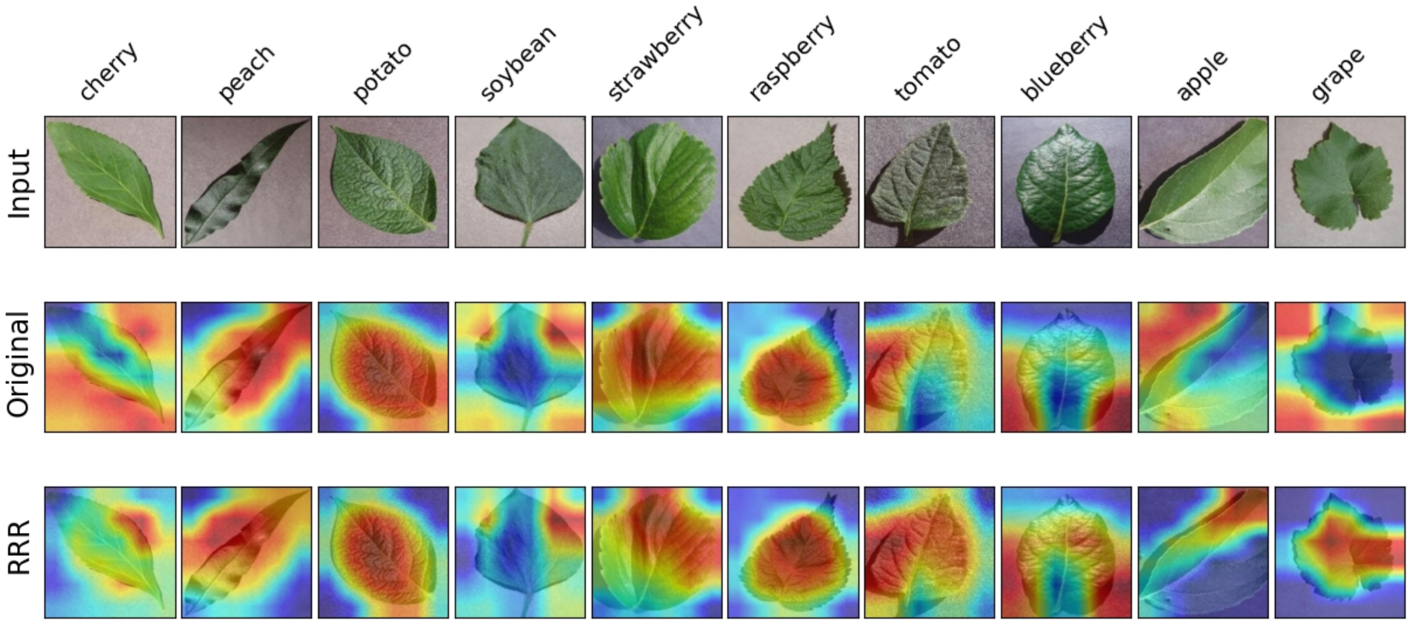
- Real masks: background pixels



#3 Classifying plant species

- Real masks: pixels outside the
- minimum bounding rectangles of leaves





Conclusions

- The introduced expertise (i.e., real masks) can guide and improve deep learning models, both in classification accuracy and explanation assessment.
- Ximeng Cheng, Ali Doosthosseini, and Julian Kunkel. Improve the deep learning models in forestry based on explanations and expertise. *Frontiers in Plant Science*, 13:902105, 2022. https://doi.org/10.3389/fpls.2022.902105
- Codes: https://github.com/GISCheng/xaiforestry

Ximeng Cheng
Applied Machine Learning Group
Fraunhofer HHI
ximeng.cheng@hhi.fraunhofer.de

Fraunhofer
HHI